Breaking Barriers in AI Inference: Lenovo ThinkSystem Servers Shine in MLPerf v4.1 - Unleashing the Power of AI-Ready Infrastructure

Article

In a world where AI is revolutionizing industries and transforming businesses, Lenovo is proud to announce that the ThinkSystem servers have broken barriers and taken the top spot in an impressive 54 out of 79 MLPerf v4.1 benchmarks! This remarkable achievement solidifies our position as a leader in the AI infrastructure market, empowering organizations to unlock the full potential of their AI initiatives.

With this groundbreaking performance, Lenovo ThinkSystem servers have demonstrated their ability to handle complex AI workloads with ease and efficiency, making them an ideal choice for organizations looking to accelerate their AI projects. Whether you're a leading financial institution, a cutting-edge tech firm, or a forward-thinking healthcare provider, Lenovo ThinkSystem servers are designed to help you stay ahead of the curve in the rapidly evolving world of AI.

Lenovo dominates GenAI benchmarks

Our ThinkSystem SR680a V3 and SR685a V3 systems have taken center stage by competing against each other in several generative AI benchmarks, showcasing the incredible power of Lenovo's server configurations. The results are impressive:

- **GPT-99 Champion:** The ThinkSystem SR680a V3 (Intel) with 8x NVIDIA H200 Tensor Core SXM GPUs, each with 141GB, took top honors, leveraging its Intel 8568Y 48-core processor and advanced memory architecture to deliver exceptional performance.
- **Llama 2 Leaderboard:** We secured victory again, this time with the Lenovo ThinkSystem SR685a V3 (AMD) with 8x NVIDIA H200 SXM GPUs with 141GB, highlighting the versatility of Lenovo's server configurations across different architectures.
- **Stable Diffusion XL Speedster:** Lenovo ThinkSystem SR680a V3 (Intel) with 8x NVIDIA H200 SXM GPUs with 141GB came out on top in this highly competitive category, demonstrating its ability to handle complex AI workloads and scale performance for demanding applications.

Outstanding MLPerf results

With finishes at or near the top of the pack across numerous MLPerf Inference tests, these results showcase the capabilities of Lenovo ThinkSystem servers in various AI inference scenarios. Whether you're developing cutting-edge AI models or processing large datasets, Lenovo's server configurations provide the scalability and performance you need to drive innovation forward.

The following table provides a breakdown of our results.

Table 1. MLPerf results

System	Total Categories Submitted	First Place Finishes	Second Place Finishes	Third Place Finishes
ThinkSystem SR675 V3	16	16	0	0
ThinkEdge SE455 V3	12	12	0	0
ThinkEdge SE360 V2	10	10	0	0
ThinkSystem SR650 V3	9	9	0	0
ThinkSystem SR680a V3 (Intel)	16	6	7	2
ThinkSystem SR685a V3 (AMD)	16	1	1	6

Our benchmarks resulted in world records against all tested systems showcasing Lenovo’s consistent improvement to achieve best-in-class results for our customers. These key results, in particular, show how powerful our systems are in those specific categories:

- **ThinkSystem SR685a V3(8x H200-SXM-141GB, NVIDIA TensorRT) - 1 st place on retinanet offline**
- **ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) - 1 st place on bert-99.9 server**
- **ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 1 st place on gptj-99 offline**
- **ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 1 st place on gptj-99.9 offline**
- **ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 1 st place on stable-diffusion-xl server**
- ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 2 nd place on stable-diffusion-xl offline
- ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) - 2 nd place on bert-99.9 server
- ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 3 rd place on gptj-99 server
- ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) - 3 rd place on 3d-unet-99.9 offline
- ThinkSystem SR680a V3 (8x H200-SXM-141GB, NVIDIA TensorRT) – 3 rd place on gptj-99.9 server

Conclusion

The insights from the latest MLPerf benchmarks are critical for stakeholders in the generative AI and machine learning ecosystem, from system architects to application developers. They provide a quantitative foundation for hardware selection and optimization, crucial for deploying scalable and efficient AI/ML systems. Future developments in hardware and software are anticipated to further influence these benchmarks, continuing the cycle of innovation and evaluation in the field of machine learning.

Professionals in the field are encouraged to consider these results in their future hardware procurement and system design strategies. For further discussion or consultation on leveraging these insights in specific use cases, engage with our expert team at aidiscover@lenovo.com.

For more information

For more information, see the following resources:

- Explore Lenovo AI solutions:
  https://www.lenovo.com/us/en/servers-storage/solutions/ai/
- Engage the Lenovo AI Center of Excellence:
  https://lenovo-ai-discover.atlassian.net/servicedesk/customer/portal/3
- MLCommons®, the open engineering consortium and leading force behind MLPerf, has now released new results for MLPerf benchmark suites:
  - Benchmark results: https://mlcommons.org/benchmarks/inference-datacenter/
  - Latest news about MLCommons: https://mlcommons.org/news-blog

Author

David Ellison is the Chief Data Scientist for Lenovo ISG. Through Lenovo's US and European AI Discover Centers, he leads a team that uses cutting-edge AI techniques to deliver solutions for external customers while internally supporting the overall AI strategy for the World Wide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the US Postal Service. Previous to that, he received a PhD in Biomedical Engineering from Johns Hopkins University. He has numerous publications in top tier journals including two in the Proceedings of the National Academy of the Sciences.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- MLPerf Benchmark

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2036, was created or updated on October 11, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2036
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2036.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkEdge®
ThinkSystem®

The following terms are trademarks of other companies:

AMD is a trademark of Advanced Micro Devices, Inc.

Intel® is a trademark of Intel Corporation or its subsidiaries.

Other company, product, or service names may be trademarks or service marks of others.